Appl. Comput. Harmon. Anal. 33 (2012) 261–281

Contents lists available at SciVerse ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

An efficient tree-based computation of a metric comparable to a natural diffusion distance

Maxim J. Goldberg [a,*], Seonja Kim [b]

[a] Theoretical and Applied Science, Ramapo College of NJ, 505 Ramapo Valley Road, Mahwah, NJ 07430, United States
[b] Department of Mathematics, CS, and Statistics, Bloomsburg University of PA, 400 E. Second St., Bloomsburg, PA 17815, United States

ARTICLE INFO

Article history:
Received 5 September 2010
Revised 27 June 2011
Accepted 12 December 2011
Available online 14 December 2011
Communicated by Charles K. Chui

Keywords:
Tree
Diffusion
Distance
Metric

ABSTRACT

Using diffusion to define distances between points on a manifold (or a sampled data set) has been successfully employed in various applications such as data organization and approximately isometric embedding of high dimensional data in low dimensional Euclidean space. Recently, P. Jones has proposed a diffusion distance which is both intuitively appealing and scales appropriately with increasing time. In the first part of our paper, we present an efficient tree-based approach to computing an approximation to Jones's diffusion distance. We also show our approximation is comparable to Jones's distance. Neither Jones's distance, nor our approximation, satisfies the triangle inequality; in particular, in the case of heat flow on $\mathbb{R}^n$, Jones's separation distance gives a scaled square of the Euclidean distance. In the second part of our paper, we present a general construction to obtain an "almost" metric from a general distance. We also discuss a numerical procedure to implement our construction. Additionally, we show that in the case of heat flow on $\mathbb{R}^n$, we recover (scaled) Euclidean distance from Jones's distance.

1. Introduction

Several years ago, motivated by considering heat flow on a manifold, R. Coifman (Yale University) proposed a diffusion metric—both for the case of a manifold and a discrete analog for a set of data points in $\mathbb{R}^n$. In the continuous case, his metric can be written as the L^2 norm of the difference of two specified vectors, each of which has unit L^1 norm. (An analogous situation holds in the discrete case.) Coifman's metric can be successfully used in various applications, including data organization, approximately isometric embedding of data in low dimensional Euclidean space, etc. See, for example, [4,5,10,13].

Neither Coifman's original distance, nor its L^1 version, namely $\|\rho_t(\cdot, x) - \rho_t(\cdot, y)\|_1$, results in (scaled) Euclidean distance for the case of heat flow on $\mathbb{R}^n$, which would be an appealing distance to have for this basic situation. As we will see in Section 13, the L^1 version is a non-linear function of the Euclidean distance divided by the square root of time, for the case of heat flow on $\mathbb{R}^n$.

For the interested reader, we present a more detailed discussion of Coifman's original distance, including what we view as some of its drawbacks, in Appendix B.

As communicated to us in a conversation with R. Coifman, P. Jones (Yale University) has proposed another notion of diffusion distance which is appropriately local, intuitively appealing, and does not use globally defined (eigenvector) functions as does Coifman's definition. (Jones's distance does not satisfy the triangle inequality, so we use the term "distance" rather than "metric"; see Sections 2 as well as 6, and 13.) As we show later in the paper, see Section 13, Jones's distance gives the

* Corresponding author.
E-mail addresses: mgoldber@ramapo.edu (M.J. Goldberg), seonja777@hotmail.com (S. Kim).

1063-5203/$ – see front matter
doi:10.1016/j.acha.2011.12.001

(scaled) *square* of the usual Euclidean distance for heat flow on $\mathbb{R}^n$ both locally and globally. Moreover, the scaling constant does not depend on the dimension n. (Coifman's original diffusion distance for heat flow on $\mathbb{R}^n$ gives Euclidean distance only locally, with a constant which depends both on the time t chosen and the dimension n; see Appendix B.) In the second part of our paper, we present a general construction for distances which turns the (scaled) square of Euclidean distance into (scaled) Euclidean distance itself. We view the fact that Jones's distance leads to the Euclidean metric for the case of heat flow on $\mathbb{R}^n$ to be a good indication that Jones's proposed distance is valuable to explore for various data sets.

The idea of Jones's proposal (see Section 3 for the details) is the following. For two points x and y in our space (say a manifold), consider a particle of unit mass diffusing from the point x, and a particle of unit mass diffusing independently from the point y. Conceptually, the initial delta function densities at x and y, respectively, are spreading out into less and less localized bump functions as time increases. At each moment of time t, we can compare the two bump functions via the L^1 norm of their difference (essentially P. Jones's original proposal), or some other measure of separation. Since the bump functions are normalized to have L^1 norm 1, L^1 is natural to use. Note that the bumps, in general, "spread out" as time passes, and thus comparing them at a time t is appropriately scaled to time t. P. Jones's suggestion is to define the diffusion distance between x and y to be the *first time* that the two bump functions are sufficiently close to one another.

Our paper consists of two largely independent, yet connected, parts.

In the first part of the paper, we present a tree-based approach to computing an approximation to Jones's diffusion distance (see Section 4). If we start with a large data set S, computing Jones's distance for all pairs of points in S would be expensive: we would need to calculate powers of the transition matrix of size N by N, where N is the number of points in the data set. A tree organization which allows us to define a "two-sided" approximation to Jones's distance (see Lemmas 4.2 and 4.3, and the discussion that follows) yields a more efficient way to evaluate how far "apart" any two points are. In our construction, stopping when ancestors of two points are neighbors, rather than being equal, has the effect of avoiding grid artifacts in defining the approximate distance and the necessity of averaging over different realizations of ancestral sequences (which would lead to complicated probabilistic arguments).

In the second part of our paper, we describe a general method of constructing an "almost" metric from a distance. Jones's proposed diffusion distance, as well as the tree-based approximation to Jones's distance constructed in the first part of the paper, are both distances. In Section 8, we specify what we mean by "almost" in the process of describing our construction: roughly, the triangle inequality is guaranteed to be satisfied only if the two points we start with are not too close to each other, and the intermediate point we pick is not too close to either one of the two initial points.

We devote a large part of our paper, the second part, to showing how an "almost" metric can be constructed from a distance. We view the triangle inequality as conceptually, mathematically, and aesthetically pleasing: to us, a fundamental underlying concept of a "good" distance is that the distance from A to B should not be greater than the sum of distances through an intermediate point C. We think this naturality and the fact that so much theory in mathematics is built using the triangle inequality is sufficient justification to desire a metric. We thus view the second part of our paper as having independent value of its own.

On a more immediate level, when we measure distances in "real" life, we most often use the usual Euclidean distance, not, say, the square of the Euclidean distance which does not satisfy the triangle inequality. In Section 13, we show that applying our construction from the first part of the paper to the case of heat flow on $\mathbb{R}^n$ yields precisely the (scaled) square of the Euclidean distance; our procedure in the second part of the paper produces (scaled) Euclidean distance itself.

In Section 10, we present figures for a synthetic data set which, in part, show sizable regions for which the triangle inequality fails for specified pairs of points (what we later refer to later as "bad" sets). As further confirmation of the importance of the triangle inequality, we suggest that the reader imagine a disease or a rumor spreading throughout the data set in that section. If one did not adjust the proposed diffusion distance to satisfy the triangle inequality (at least for most points), one would have the counterintuitive situation that a disease/rumor spreading from point A to point B may take *more* time than spreading from A to B through an intermediate point C.

The organization of our paper is as follows. After a section on notation and assumptions for the first part of the paper, we describe Jones's proposed distance and give some equivalent formulations in Section 3. In the following section, Section 4, we describe our tree construction to approximate Jones's distance. We view this section as the main part of the paper. In Section 5, we illustrate the construction described in Section 4 for a synthetic data set.

Section 6 presents the notation and assumptions for the second part of the paper: constructing an "almost" metric from a distance. In Section 7 we illustrate our underlying definitions for $\mathbb{R}^n$ and the scaled square of Euclidean distance. Section 8 follows, and gives the general construction. In Section 9, we illustrate our construction for an example of a distance in $\mathbb{R}^n$, while in Section 10 we return to our synthetic data set introduced in Section 5. In the following section (Section 11) we outline a numerical procedure to implement our general construction. After a brief section which discusses our metric construction in relation to Jones's distance and the tree approximation to Jones's distance from the first part of the paper, Section 13 applies our construction to Jones's distance for heat flow in $\mathbb{R}^n$. In Section 14, we introduce a variation of Jones's distance and make some remarks about the applicability of our construction of a metric to more general cases than heat flow in $\mathbb{R}^n$. We conclude the paper with a conclusions section, Appendix A consisting of some computations, Appendix B which discusses Coifman's diffusion distance in more detail, and acknowledgments.

2. Notation and assumptions for the first part of the paper

We list some notations and assumptions that will be used mainly through Section 4; some of the following will also appear in the rest of the paper.

Let X be a topological space equipped with a measure. For $t \geqslant 0$, $\rho_t(x, y)$ will denote a kernel on $X \times X$, with $\rho_t(x, y) \geqslant 0$ for all $x, y \in X$. We assume that ρ satisfies the semi-group property:

$$\int_X \rho_t(x, u)\rho_s(u, y)\, du = \rho_{t+s}(x, y), \tag{1}$$

for all $x, y \in X$, and $s, t \geqslant 0$. In addition, we assume the following property:

$$\int_X \rho_t(x, y)\, dx = 1, \tag{2}$$

for all $y \in X$ and all $t \geqslant 0$. The latter convention gives the mass preservation property (for f in an appropriate functional or measure space):

$$\int_X T_t f(x)\, dx = \int_X f(y)\, dy, \tag{3}$$

where

$$T_t f(x) \equiv \int_X \rho_t(x, y) f(y)\, dy. \tag{4}$$

We assume $\rho_0(x, u) = \delta_x(u)$, where $\delta_x(u)$ denotes the Dirac delta function centered at x. We do NOT necessarily assume that $\rho_t(x, y)$ is symmetric in its arguments, nor that the integral with respect to the second variable is 1.

We will refer to a kernel ρ_t satisfying the various conditions above as a diffusion kernel (at time t); T_t is the diffusion operator at time t. A typical example for ρ_t is the heat kernel on a Riemannian manifold (the heat kernel is symmetric, see [3], but we do not assume symmetry of ρ_t in general).

By the mass preservation property, Eq. (3) above, for every $t > 0$, we see that T_t can be viewed as an operator from the convex space of probability measures on the space X into itself. We will assume that

$$\left\| T_t(\delta_{y_1} - \delta_{y_2}) \right\|_1 = \left\| \rho_t(\cdot, y_1) - \rho_t(\cdot, y_2) \right\|_1 \to 0, \quad \text{as } t \to \infty, \tag{5}$$

for any points y_1 and y_2. (For related discussion about existence and uniqueness of a fixed point for T_t, see for example [9] and [14].)

In what follows, we will consider functions $\tau : X \times X \to [0, \infty)$ with the following properties:

$$\tau(x, y) = 0 \iff x = y; \qquad \tau(x, y) = \tau(y, x). \tag{6}$$

In [7], such a function τ satisfying a slightly weaker requirement, namely that τ is non-negative, symmetric, and $\tau(x, x) = 0$ (instead of the "if and only if requirement" in Eq. (6)) is called a distance. We will call our τ a distance as well. Note we are not assuming the triangle inequality.

We would like to note that there are many notions to measure separation between points, with various authors sometimes differing as to what terminology refers to which properties. Chapter 1 of [7] provides a compendium of more than 20 measures of separation, including distance, similarity, semi-metric, metric, extended metric, near-metric, quasi-distance, quasi-semi-metric, Albert quasi-metric, weak quasi-metric, quasi-metric, etc. In the present paper we are concentrating on only two such measures of separation: distance as we discuss just above, and metric, i.e., a distance for which the triangle inequality holds (see also Section 6). We thus do not feel it necessary to introduce or discuss other measures of separation and suggest that the interested reader consult [7], for example.

Finally, $\| \ \|$ will denote the norm $1/2\| \ \|_1$, one-half the L^1 norm.

3. A distance proposed by Peter Jones

In this section, we present Peter Jones's proposal for the diffusion distance between two points. We also discuss some easy reformulations of his definition.

Peter Jones (private communication) has proposed the following as a measure of the distance between two points x and y in X: let $\tau_J(x, y)$ equal the greatest lower bound of the times t so that $\int_X \min(\rho_t(u, x), \rho_t(u, y))\, du \geqslant \gamma$, for a fixed parameter γ with $0 < \gamma < 1$, i.e.,

$$\tau_J(x,y) = \operatorname{arginf}_t \left\{ \int_X \min\bigl(\rho_t(u,x), \rho_t(u,y)\bigr)\, du \geqslant \gamma \right\}. \tag{7}$$

The definition of τ_J depends on γ and thus, strictly speaking, we should include the dependence on γ in the notation explicitly. For ease of notation, we have omitted denoting γ explicitly.

Note that τ_J does satisfy the requirement for being a distance as listed in Section 2, namely:

$$\tau_J(x,y) = 0 \quad \Longleftrightarrow \quad x = y; \qquad \tau_J(x,y) = \tau_J(y,x). \tag{8}$$

The first property follows since ρ_0 is assumed to be the delta function, and the second property from the definition of τ. We emphasize that the symmetry in x and y holds even if ρ_t is not symmetric in its arguments.

We observe that:

$$\begin{aligned}
\bigl\|\rho_t(\cdot,x) - \rho_t(\cdot,y)\bigr\| &\equiv \frac{1}{2}\bigl\|\rho_t(\cdot,x) - \rho_t(\cdot,y)\bigr\|_1 \\
&= \frac{1}{2}\int_X \bigl|\rho_t(u,x) - \rho_t(u,y)\bigr|\, du \\
&= \frac{1}{2}\biggl(\int_X \max\bigl(\rho_t(u,x), \rho_t(u,y)\bigr) - \min\bigl(\rho_t(u,x), \rho_t(u,y)\bigr)\, du\biggr) \\
&= \frac{1}{2}\biggl(\int_X \max\bigl(\rho_t(u,x), \rho_t(u,y)\bigr) + \min\bigl(\rho_t(u,x), \rho_t(u,y)\bigr) - 2\min\bigl(\rho_t(u,x), \rho_t(u,y)\bigr)\, du\biggr) \\
&= \frac{1}{2}\biggl(\int_X \rho_t(u,x)\, du + \int_X \rho_t(u,y)\, du - 2\int_X \min\bigl(\rho_t(u,x), \rho_t(u,y)\bigr)\, du\biggr) \\
&= 1 - \int_X \min\bigl(\rho_t(u,x), \rho_t(u,y)\bigr)\, du.
\end{aligned} \tag{9}$$

This observation is well-known, see for example entry 11 in Table 3 in [2]. In addition, defining probability measures $\mu_{t,x}(A)$ and $\nu_{t,y}(A)$ by $\mu_{t,x}(A) = \int_A \rho_t(u,x)\, du$ and $\nu_{t,y}(A) = \int_A \rho_t(u,y)\, du$ for measurable $A \subseteq X$, we also have that

$$\frac{1}{2}\bigl\|\rho_t(\cdot,x) - \rho_t(\cdot,y)\bigr\|_1 = \|\mu_{t,x} - \nu_{t,y}\|_{\mathrm{TV}} \equiv \max_A \bigl|\mu_{t,x}(A) - \nu_{t,y}(A)\bigr|, \tag{10}$$

where the maximum is taken over all measurable $A \subseteq X$, see Chapter 4 of [11].

The following results are easy to establish:

Proposition 3.1. $\|T_t f\|_1 \leqslant \|f\|_1$.

Proof. Using Fubini, we have:

$$\begin{aligned}
\int \bigl|T_t f(x)\bigr|\, dx &\leqslant \iint \rho_t(x,y)\bigl|f(y)\bigr|\, dy\, dx \\
&= \int \biggl(\int \rho_t(x,y)\, dx\biggr)\bigl|f(y)\bigr|\, dy \\
&= \int \bigl|f(y)\bigr|\, dy = \|f\|_1. \qquad \square
\end{aligned} \tag{11}$$

Corollary 3.1. *$\|T_t f\|_1$ is decreasing in t.*

Proof. By the above proposition,

$$\|T_{t+s} f\|_1 = \bigl\|T_s(T_t f)\bigr\|_1 \leqslant \|T_t f\|_1. \qquad \square \tag{12}$$

Proposition 3.2.

$$\bigl\|\rho_t(\cdot,y_1) - \rho_t(\cdot,y_2)\bigr\| = \frac{1}{2}\bigl\|T_t(\delta_{y_1} - \delta_{y_2})\bigr\|_1 \leqslant \frac{1}{2}\|\delta_{y_1} - \delta_{y_2}\|_1 = 1, \tag{13}$$

for any $y_1 \neq y_2$. Hence, $\|\rho_t(\cdot,y_1) - \rho_t(\cdot,y_2)\|$ is 1 when $t = 0$, and decreases as $t \to \infty$.

Combining Eq. (9) and Corollary 3.1, we see that Jones's distance $\tau_J(x, y)$ is:

$$\begin{aligned}\tau_J(x,y) &= \operatorname{arginf}_t\left\{\int_X \min\bigl(\rho_t(u,x),\rho_t(u,y)\bigr)\,du \geqslant \gamma\right\}\\ &= \operatorname{arginf}_t\left\{\inf_{s\geqslant t}\left\{\int_X \min\bigl(\rho_s(u,x),\rho_s(u,y)\bigr)\,du\right\} \geqslant \gamma\right\}\\ &= \operatorname{arginf}_t\bigl\{\bigl\|\rho_t(\cdot,x)-\rho_t(\cdot,y)\bigr\| \leqslant 1-\gamma\bigr\}\\ &= \operatorname{arginf}_t\Bigl\{\sup_{s\geqslant t}\bigl\{\bigl\|\rho_s(\cdot,x)-\rho_s(\cdot,y)\bigr\|\bigr\} \leqslant 1-\gamma\Bigr\}. \end{aligned} \tag{14}$$

Note that in view of Proposition 3.2, finding Jones's distance between the points x and y amounts to inverting the function $t \longrightarrow \|\rho_t(\cdot, x) - \rho_t(\cdot, y)\|$.

4. A tree construction to approximate Jones's distance

In this section, which we view as the main contribution of our paper, we discuss a tree construction which allows us to efficiently compute an approximation to Jones's distance.

We will discuss our construction for a discrete data set, S. We will mix continuous and discrete notation freely; it should be clear from the context whether integration or summation is to be used. The ω which appears in this section equals $1-\gamma$ in (14) above.

When we make the assumption that the cost of finding the diffusion neighbors of any point in a subset of the original data set S (with the property that any two elements of the subset are comparably separated relative to the neighborhood size) is bounded by a constant, we show that the computational cost of our construction is bounded by a constant times the size of S. (See Assumption 4.1, pseudocode, and the attendant discussion near the end of this section.)

For the convenience of the reader, we now summarize our algorithm which follows. Our inputs are a discrete data set S, a row stochastic transition matrix A prescribing a random walk on S, and a number $\omega > 0$, which will be a measure of diffusion proximity. Our output consists of a sequence of sets S_j which give successively coarser partitions of S (see below) and a set of "ancestor" points for every point $y \in S$ (with the jth ancestor in S_j). For any $y_1, y_2 \in S$, we find the first level n so that nth level ancestor of y_1 and the nth level ancestor of y_2, are neighbors as defined below (see inequality (20)). Finally, we define the approximation to Jones's distance between y_1 and y_2 to be 2^n. Lemmas 4.2 and 4.3 show how our approximation compares to Jones's distance between y_1 and y_2 from above and from below. Stopping when ancestors of two points are neighbors, rather than being equal, has the effect of avoiding grid artifacts in defining the approximate distance and the necessity of averaging over different realizations of ancestral sequences (which would lead to complicated probabilistic arguments).

Our algorithm as described is implemented for a specific synthetic data set in Section 5, exactly as we now describe it.

Suppose that ρ_1, i.e., the kernel above at $t = 1$, is specified by a row stochastic transition matrix A; more explicitly,

$$A(y, x) \equiv \rho_1(x, y). \tag{15}$$

In the case of heat flow, A would be the discretization of the heat kernel on a discrete data set, at unit time. As mentioned in Section 2, the heat kernel is symmetric and thus A would be symmetric as well. However, in important applications, A need not arise from heat flow and need not be symmetric. A model example we are thinking about is the following (this example is mentioned in [8]). Suppose we have a map grid, and are tracking some localized storm which is currently at some particular location on the grid. We suppose that the storm behaves like a random walk, and has a certain (constant in time) probability to move from one grid location to another at each "tick of the clock" (time step). We can thus model the movements of the storm by a Markov matrix A, with the nth power of A giving the transition probabilities after n ticks of the clock. If there is no overall wind, the matrix A could reasonably be assumed to be symmetric. But suppose there is an overall wind in some fixed direction, which is making it more probable for the storm to move north, say, rather than south. Then the matrix A is not symmetric, there is a preferred direction of the storm to move in, from one tick of the clock to the next.

Assume that $A^{1,000,000}$ is a high enough power to start having data points interacting with nearby points. For instance, this should happen if the second eigenvalue of A is, say, less than $1 - 10^{-6}$. Raising $A^{1,000,000}$ to an additional 1,000,000th power, i.e., raising the original transition matrix A to the power 10^{12}, should reach the limiting measure in every row to any reasonable precision. Thus in what follows, when we consider A^{2^n}, n will be less than or equal to 40 (since $2^{40} > 10^{12}$).

Using our discussion in the previous section, see Proposition 3.2, assuming (5), see Section 2, we have that $1/2\|\rho_t(\cdot, y_1) - \rho_t(\cdot, y_2)\|_1$ is 1 when $t = 0$ (if $y_1 \neq y_2$), and decreases to 0 as $t \to \infty$.

Now, choose $\omega > 0$, which will be a measure of diffusion proximity. We assume that $\omega/41$ is such that $A^{2^{40}}$, i.e., $t = 2^{40}$, is "time enough" to bring diffusions originating at any two points of S within $\omega/41$ of each other, in the $\|\ \| = 1/2\|\ \|_1$ norm. We think this is a reasonable assumption.

We then build a sequence of sets $S_{-1}, S_0, S_1, \ldots, S_{40}$ such that the following hold. We have that

$$S = S_{-1} \supseteq S_0 \supseteq S_1 \supseteq S_2 \supseteq \cdots \supseteq S_{40}, \tag{16}$$

with S_{40} consisting of just one point. Furthermore, for $-1 \leqslant n \leqslant 40$,

$$S_n = \{y_1^{(n)}, y_2^{(n)}, \ldots, y_{i_n}^{(n)}\} \tag{17}$$

(where $y_i^{(-1)} \equiv y_i \in S$) and with

$$\|\rho_{2^n}(\cdot, y_i^{(n)}) - \rho_{2^n}(\cdot, y_j^{(n)})\| \geqslant \omega/41, \quad i \neq j, \tag{18}$$

and, if $n \geqslant 0$, for every $y_k^{(n-1)} \in S_{n-1}$, $\exists y_i^{(n)} \in S_n$ with

$$\|\rho_{2^n}(\cdot, y_k^{(n-1)}) - \rho_{2^n}(\cdot, y_i^{(n)})\| < \omega/41. \tag{19}$$

Note that the $y_i^{(n)}$ in the previous equation need not be unique.

At each level n, for each $y_i^{(n)} \in S_n$, we keep track of the *neighbors* of $y_i^{(n)}$, i.e., those points $y_j^{(n)} \in S_n$ such that

$$\|\rho_{2^n}(\cdot, y_i^{(n)}) - \rho_{2^n}(\cdot, y_j^{(n)})\| < 3\omega. \tag{20}$$

Note: ω, not $\omega/41$, is used, so $3\omega = 123 \times \omega/41$, giving a rather large neighborhood relative to $\omega/41$, but still doable.

Note 4.1. We are only keeping track of sets of points, the S_n's, at each level; we won't need to consider any associated covers of S.

Now, to every $y \in S$ associate its "ancestor" sequence:

$$y_{i(y)}^{(0)}, y_{i(y)}^{(1)}, \ldots, y_{i(y)}^{(40)}, \tag{21}$$

where $y_{i(y)}^{(0)} \in S_0$ and

$$\|\rho_1(\cdot, y) - \rho_1(\cdot, y_{i(y)}^{(0)})\| < \omega/41, \tag{22}$$

and, for $n = 1, \ldots, 40$, $y_{i(y)}^{(n)} \in S_n$ and

$$\|\rho_{2^n}(\cdot, y_{i(y)}^{(n-1)}) - \rho_{2^n}(\cdot, y_{i(y)}^{(n)})\| < \omega/41. \tag{23}$$

Of course, more than one such ancestral sequence may exist due to the possibility of several choice of $y_{i(y)}^{(n)}$ after $y_{i(y)}^{(n-1)}$ is selected. We just keep track of just ONE possible ancestral sequence. For $n = 0, 1, \ldots, 40$, we will call $y_{i(y)}^{(n)}$ (in the chosen ancestral sequence of y) the nth ancestor of y.

As pointed out by one of the reviewers, the notation $y_{i(y)}^{(n)}$ is improper since the index $i(y)$ depends not only on y but also on the level n. In the above, and in what follows, we will keep the notation $y_{i(y)}^{(n)}$ for simplicity, and remind the reader that it is simplified notation for $y_{i(y;n)}^{(n)}$.

The following simple lemma is important in connecting the diffusion from a point $y \in S$ to its nth ancestor at time 2^n.

Lemma 4.1. *For $n = 0, 1, \ldots, 40$, if $y_{i(y)}^{(n)}$ is the nth ancestor of a point y in the data set S, we have*:

$$\|\rho_{2^n}(\cdot, y) - \rho_{2^n}(\cdot, y_{i(y)}^{(n)})\| < \frac{n+1}{41}\omega. \tag{24}$$

Proof. The proof is a simple application of the triangle inequality and decrease in t of $\|\rho_t(\cdot, y_1) - \rho_t(\cdot, y_2)\|$ (see Proposition 3.2). By the triangle inequality for L^1 (remembering that $\| \ \| = 1/2\| \ \|_1$), we have:

$$\begin{aligned}
\|\rho_{2^n}(\cdot, y) - \rho_{2^n}(\cdot, y_{i(y)}^{(n)})\| \leqslant{}& \|\rho_{2^n}(\cdot, y) - \rho_{2^n}(\cdot, y_{i(y)}^{(0)})\| \\
&+ \|\rho_{2^n}(\cdot, y_{i(y)}^{(0)}) - \rho_{2^n}(\cdot, y_{i(y)}^{(1)})\| \\
&+ \cdots \\
&+ \|\rho_{2^n}(\cdot, y_{i(y)}^{(n-2)}) - \rho_{2^n}(\cdot, y_{i(y)}^{(n-1)})\| \\
&+ \|\rho_{2^n}(\cdot, y_{i(y)}^{(n-1)}) - \rho_{2^n}(\cdot, y_{i(y)}^{(n)})\|.
\end{aligned} \tag{25}$$

By the decrease in t of $\|\rho_t(\cdot, y_1) - \rho_t(\cdot, y_2)\|$ noted above, and inequalities (22) and (23), we have:

$$\begin{aligned}
&\left\|\rho_{2^n}(\cdot, y) - \rho_{2^n}\left(\cdot, y^{(0)}_{i(y)}\right)\right\| \leqslant \left\|\rho_1(\cdot, y) - \rho_1\left(\cdot, y^{(0)}_{i(y)}\right)\right\| < \omega/41,\\
&\left\|\rho_{2^n}\left(\cdot, y^{(0)}_{i(y)}\right) - \rho_{2^n}\left(\cdot, y^{(1)}_{i(y)}\right)\right\| \leqslant \left\|\rho_2\left(\cdot, y^{(0)}_{i(y)}\right) - \rho_2\left(\cdot, y^{(1)}_{i(y)}\right)\right\| < \omega/41,\\
&\quad\vdots\\
&\left\|\rho_{2^n}\left(\cdot, y^{(n-2)}_{i(y)}\right) - \rho_{2^n}\left(\cdot, y^{(n-1)}_{i(y)}\right)\right\| \leqslant \left\|\rho_{2^{n-1}}\left(\cdot, y^{(n-2)}_{i(y)}\right) - \rho_{2^{n-1}}\left(\cdot, y^{(n-1)}_{i(y)}\right)\right\| < \omega/41, \quad \text{and}\\
&\left\|\rho_{2^n}\left(\cdot, y^{(n-1)}_{i(y)}\right) - \rho_{2^n}\left(\cdot, y^{(n)}_{i(y)}\right)\right\| < \omega/41. \qquad (26)
\end{aligned}$$

Combining (25) and (26), we obtain:

$$\left\|\rho_{2^n}(\cdot, y) - \rho_{2^n}\left(\cdot, y^{(n)}_{i(y)}\right)\right\| < \frac{n+1}{41}\omega. \qquad \square \qquad (27)$$

Now, choose any $y_1, y_2 \in S$. Let n be the FIRST level n so that $y^{(n)}_{i(y_1)}$, the nth ancestor of y_1, and $y^{(n)}_{i(y_2)}$, the nth ancestor of y_2, are neighbors, see inequality (20).

Lemma 4.2.

$$\left\|\rho_{2^n}(\cdot, y_1) - \rho_{2^n}(\cdot, y_2)\right\| \leqslant 3\omega + 2\frac{n+1}{41}\omega. \qquad (28)$$

Proof. We see that:

$$\begin{aligned}
\left\|\rho_{2^n}(\cdot, y_1) - \rho_{2^n}(\cdot, y_2)\right\| &\leqslant \left\|\rho_{2^n}(\cdot, y_1) - \rho_{2^n}\left(\cdot, y^{(n)}_{i(y_1)}\right)\right\|\\
&\quad + \left\|\rho_{2^n}\left(\cdot, y^{(n)}_{i(y_1)}\right) - \rho_{2^n}\left(\cdot, y^{(n)}_{i(y_2)}\right)\right\|\\
&\quad + \left\|\rho_{2^n}\left(\cdot, y^{(n)}_{i(y_2)}\right) - \rho_{2^n}(\cdot, y_2)\right\|\\
&< \frac{n+1}{41}\omega + 3\omega + \frac{n+1}{41}\omega = 3\omega + 2\frac{n+1}{41}\omega. \qquad (29)
\end{aligned}$$

We have used Lemma 4.1 and that $y^{(n)}_{i(y_1)}$ and $y^{(n)}_{i(y_2)}$ (the nth ancestors of, respectively, y_1, y_2) are neighbors. $\square$

Lemma 4.3. *If $n \geqslant 1$, we have*:

$$\left\|\rho_{2^{n-1}}(\cdot, y_1) - \rho_{2^{n-1}}(\cdot, y_2)\right\| \geqslant 3\omega - 2\frac{n}{41}\omega. \qquad (30)$$

Proof. We see that:

$$\begin{aligned}
\left\|\rho_{2^{n-1}}(\cdot, y_1) - \rho_{2^{n-1}}(\cdot, y_2)\right\| &\geqslant \left\|\rho_{2^{n-1}}\left(\cdot, y^{(n-1)}_{i(y_1)}\right) - \rho_{2^{n-1}}\left(\cdot, y^{(n-1)}_{i(y_2)}\right)\right\|\\
&\quad - \left\|\rho_{2^{n-1}}\left(\cdot, y^{(n-1)}_{i(y_1)}\right) - \rho_{2^{n-1}}(\cdot, y_1)\right\|\\
&\quad - \left\|\rho_{2^{n-1}}\left(\cdot, y^{(n-1)}_{i(y_2)}\right) - \rho_{2^{n-1}}(\cdot, y_2)\right\|\\
&\geqslant 3\omega - \frac{n}{41}\omega - \frac{n}{41}\omega = 3\omega - 2\frac{n}{41}\omega, \qquad (31)
\end{aligned}$$

using Lemma 4.1 and that $y^{(n-1)}_{i(y_1)}$ and $y^{(n-1)}_{i(y_2)}$ are not neighbors. $\square$

Finally, for $y_1, y_2 \in S$, $y_1 \neq y_2$, we define $\tau_{approx}(y_1, y_2) = 2^n$, where n is the first level n so that $y^{(n)}_{i(y_1)}$, the nth ancestor of y_1, and $y^{(n)}_{i(y_2)}$, the nth ancestor of y_2, are neighbors. If $y_1 = y_2$, we define $\tau_{approx}(y_1, y_2) = 0$. Clearly τ_{approx} is a distance as defined in Section 2. We see that τ_{approx} is comparable to τ_J in the sense of Lemmas 4.2 and 4.3.

We conclude this section with some pseudocode for our construction. The main code fragment that we will give below uses the following subroutine:

```
Subroutine Neighborhood(X, x, t, α)
   Input: X, a finite set of points; x ∈ X; t ⩾ 0, α > 0
   Output: The set of neighbors of x at time t and closeness α, i.e.,
      {z ∈ X: ‖ρ_t(·, z) − ρ_t(·, x)‖ < α}
```

In the case of a (discrete) random walk, we would look at the probability distribution of the location of a traveler starting at a point x after (integer) time t, and compare it to the probability distribution of the location of a traveler starting at a point z after time t: if the two distributions are within α of each other in (scaled) l_1 norm, z will be called a neighbor of x.

Our assumption about out data set S and the kernel ρ is the following.

Assumption 4.1. Let r, c_1 and c_2 be constants such that $0 < r$ and $0 < c_1, c_2 < 1$. Let t_1 and t_2 be any positive times so that $t_2/t_1 \leqslant r$, and let α_1 and α_2 be positive numbers satisfying $c_1 \leqslant \alpha_1$ and $\alpha_2 \leqslant c_2$. Suppose that $\tilde{S}$ is any subset of S such that for any two distinct points $s_1, s_2 \in \tilde{S}$, we have $\|\rho_{t_1}(\cdot, s_1) - \rho_{t_1}(\cdot, s_2)\| \geqslant \alpha_1$. Then we assume that for every $s \in \tilde{S}$, the computational cost of $Neighborhood(\tilde{S}, s, t_2, \alpha_2)$ is bounded by a constant which depends only on r, c_1 and c_2. Additionally, we make the "starting" assumption that for every point in the initial data set S and every $s \in S$, the computational cost of $Neighborhood(S, s, 1, \alpha)$ is bounded by a constant for any positive α that is sufficiently small.

Loosely speaking, our assumption is that for any subset of S which has points separated at time t_1 by at least α_1 (relative to the norm $\|\ \|$ of the difference of the respective probability densities), finding all the neighbors of any point of that subset at time t_2 which are at most α_2 away has computational cost bounded by a constant (provided t_2 is not too much larger than t_1, α_1 is bounded away from 0, and α_2 is bounded away from 1). (The "starting" assumption is that at the initial time $t = 1$, the computational cost of finding the diffusion neighbors of any point s which are not too far away is also bounded by a constant.) As can easily be seen from Section 13, for the case of heat flow on $\mathbb{R}^n$ the assumption above is saying that for a subset of our data set all of whose points are at least some specified positive Euclidean distance apart from each other, the computational effort to find all points (in that subset) which are a comparable specified Euclidean distance away from a point in the subset, is bounded by a constant.

The pseudocode below constructs the sets S_n, ancestral sequences of every point in S (by successively identifying a parent at each level), and the 3ω neighbors of each element in every S_n, $n \geqslant 0$:

```
n := −1;
S_n := S;
while S_n has more than element
   n := n + 1;
   S_n := ∅;
   T := S_{n−1};
   loop over elements y of T while T ≠ ∅
      S_n := S_n ∪ {y};
      Y := Neighborhood(S_{n−1}, y, 2^n, ω/41);
      loop over x ∈ Y
         if x has not yet been assigned a parent
            parent of x := y;
         end if
      end loop
      T := T \ Y;
   end loop
   loop over all elements z ∈ S_n
      store Neighborhood(S_n, z, 2^n, 3ω);
   end loop
end while
```

By Assumption 4.1, the computational cost of each call

$$Y := Neighborhood\big(S_{n-1}, y, 2^n, \omega/41\big)$$

in the code above is bounded by a constant (using $r = 2$, $\alpha_1 = \omega/41$ and $\alpha_2 = \omega/41$), as is the cost of each call

$$\text{store } Neighborhood\big(S_n, z, 2^n, 3\omega\big)$$

(using $r = 1$, $\alpha_1 = \omega/41$ and $\alpha_2 = 3\omega$). It thus follows that the cost of executing the while loop for S_n is bounded by a constant times the size of S_n. Since the size of each S_n is bounded by the size of S, the total computational cost is bounded by CkN, where C is some constant, k is the number of times the while loop runs, and N is the size of S. In the earlier discussion in this section, we had assumed that the number of S_n's is bounded by 40, hence the cost of the computation is $O(N)$.

Once the program above is run, finding the distance between any pair of points has cost bounded by a constant: from each point, travel "up" its ancestral sequence, and stop at the first level n such that the respective nth level ancestors are 3ω neighbors in S_n. The τ_{approx} distance between these two points is then 2^n.

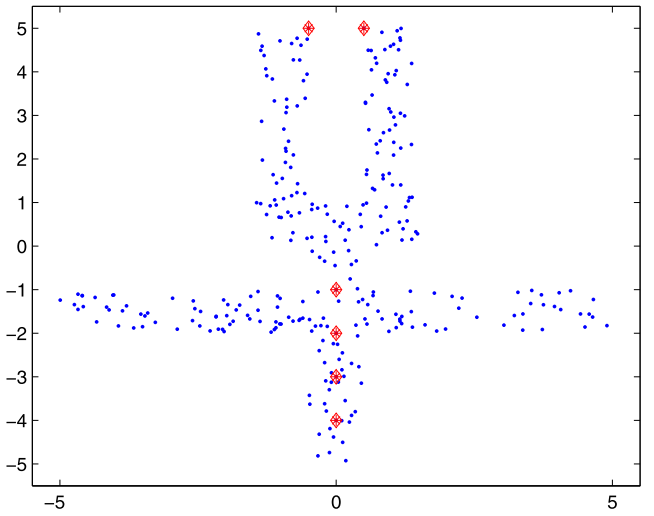

Fig. 1. The "flower" data set, the 6 additional points marked.

Table 1
τ_{approx} distances between 6 points in flower.

	$(0,-4)$	$(0,-3)$	$(0,-2)$	$(0,-1)$	$(-1/2,5)$
$(0,-3)$	8				
$(0,-2)$	32	16			
$(0,-1)$	64	32	8		
$(-1/2,5)$	512	512	512	256	
$(1/2,5)$	512	512	512	256	256

5. A numerical example

We illustrate the construction described in the previous section for a synthetic data set, which we will refer to as the "flower" data set. This set consists of 300 randomly selected points in a flower-like figure, with 6 additional points added: $(0,-4)$, $(0,-3)$, $(0,-2)$, $(0,-1)$, $(-1/2,5)$, and $(1/2,5)$. We will use these 6 points for various illustrations in what follows. Please see Fig. 1.

As we will see later in Section 13, Jones's distance for the (continuous) example of heat flow on $\mathbb{R}^n$, gives the scaled square of Euclidean distance. Although our "flower" data set lies in $\mathbb{R}^2$, we have constructed our set to have pairs of points that are close to one another in the Euclidean sense but far apart if one considers paths (which lie inside the set) connecting the points. For example, the points $(-1/2,5)$ and $(1/2,5)$, while close in the Euclidean metric, should be (and indeed turn out to be) far apart relative to τ_{approx}: a random walker starting at one of these points would "wander" for quite some time before reaching the other point. In terms of Jones's distance, a long time will pass for the two probability bumps, each spreading out from its respective point, to overlap to a significant extent. The "leaf" going across the flower's "stem" was added to see if the presence of more paths between points $(0,-2)$ and $(0,-1)$—i.e., the existence of more ways for the respective probability densities to spread out—increases the τ_{approx} distance between those points, as compared to the τ_{approx} distance between points $(0,-4)$ and $(0,-3)$.

We have implemented the construction described in Section 4 above for the flower set. For every point y in the flower set, we considered all points in the flower with (Euclidean) distance less than 0.7 from y to be in the immediate neighborhood of y, and constructed the transition matrix A as follows. For any point z in the flower set with $|y-z|<0.7$, the transition probability to go from y to z was assigned to be proportional to $e^{-|y-z|/0.7}$. The value of ω, as used in Section 4, was selected to be 0.5. (We again point out that ω in Section 4 is $1-\gamma$ in (14) in Section 3.)

Table 1 gives the τ_{approx} distances between 6 points in the data set (since τ_{approx} is symmetric, our table has a triangular structure). Note, for instance, that points $(-1/2,5)$ and $(1/2,5)$ are indeed far apart relative to the pair $(0,-4)$ and $(0,-3)$, while the two pairs of points are the same distance apart in the Euclidean metric. The presence of the leaf does not seem to increase the τ_{approx} distance, as can be seen by looking at the τ_{approx} distance between the pair of points $(0,-4)$ and $(0,-3)$ and the distance between $(0,-2)$ and $(0,-1)$. We would also like to point out that due to the dyadic nature of the construction of τ_{approx}, a factor of 2 difference when comparing distances between various pairs is not significant. In any case, in many applications determining a natural distance between two points up to some globally bounded factor is enough.

Figs. 2, 3, and 4 show the sets S_9, S_{11}, and S_{12}, respectively, for the flower set. The set S_{13} consists of one point, so the sequence of S_n's terminates with $n=13$. (For the flower set, the S_n's with $n<9$ each comprises such a large part of the original set that showing them does not seem to us to be very instructive.)

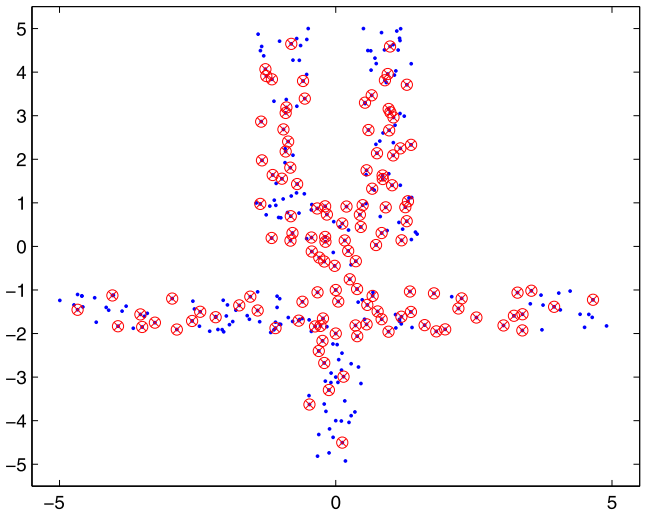

Fig. 2. The "flower" data set, S_9 marked.

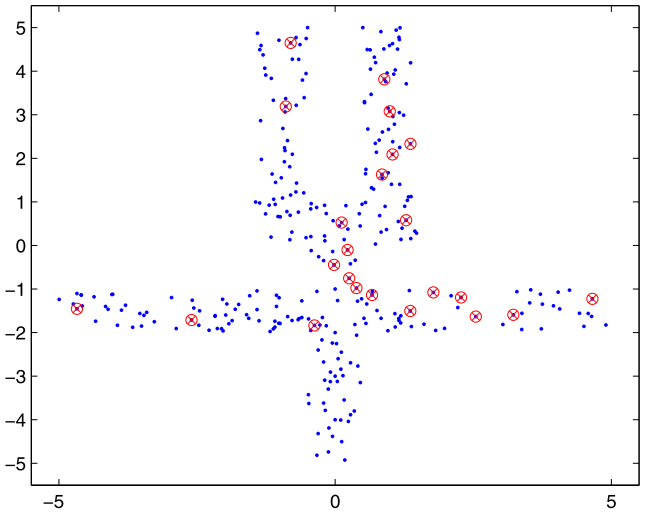

Fig. 3. The "flower" data set, S_{11} marked.

6. Notation and assumptions for the second part of the paper

In this section, we list the notation and assumptions which will be used in the second part of our paper. In what follows, X will denote a topological space (not necessarily equipped with a measure), and, as in Section 2, $\tau : X \times X \to [0, \infty)$ will be called a distance if the following properties hold:

$$\tau(x, y) = 0 \quad \Longleftrightarrow \quad x = y; \qquad \tau(x, y) = \tau(y, x). \tag{32}$$

We further assume that τ is continuous.

For $u \in X$, we define $\tau_u : X \to [0, \infty)$ by $\tau_u(z) \equiv \tau(u, z) = \tau(z, u)$. For $a \geqslant 0$, and $u \in X$, let $N_u(a) = \{z\colon\ \tau_u(z) \leqslant a\} = \tau_u^{-1}([0, a])$. For $x, y \in X$, $x \neq y$, let $B_{x,y} = \{z\colon\ \tau(x, y) \geqslant \tau(x, z) + \tau(z, y)\}$ be the "bad" set for x, y, i.e., the set where the triangle inequality fails.

Our assumptions are as follows: (1) X is such that compact $\Longrightarrow$ closed, i.e., compact sets are closed in X (this is not a stringent requirement on the topology of X); (2) for every $a \geqslant 0$, and $u \in X$, $N_u(a)$ is compact; and (3) the function τ is continuous on $X \times X$.

Note the following: $B_{x,y} \subseteq N_x(\tau(x, y)) \cap N_y(\tau(x, y))$. Since $N_x(\tau(x, y))$ and $N_y(\tau(x, y))$ are compact, hence closed, their intersection is closed, and since a closed subset of a compact set is compact, their intersection is compact. Additionally, since the functions $\tau_x(\cdot)$ and $\tau_y(\cdot)$ are continuous, and hence so is their sum, $B_{x,y} = [\tau_x + \tau_y]^{-1}([0, \tau(x, y)])$ is the inverse image of a closed set, hence also closed. Since $B_{x,y}$ is a closed subset of a compact set, $B_{x,y}$ is compact.

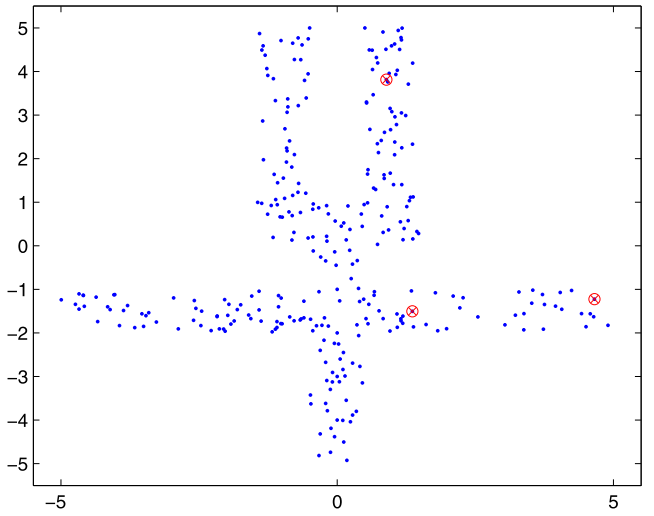

Fig. 4. The "flower" data set, S_{12} marked.

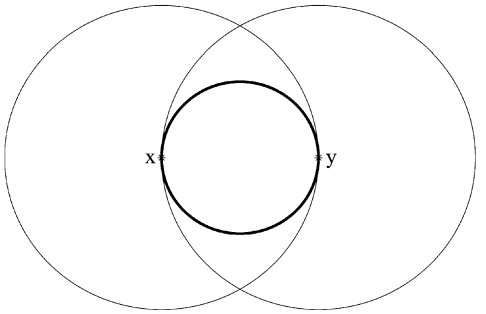

Fig. 5. The "bad" set in $\mathbb{R}^n$ for the square of the Euclidean distance.

Our goal for this part of the paper is to describe how to construct a metric out of τ, actually an "almost" metric, $d(x, y)$. Recall that $d : X \times X \to [0, \infty)$ is a metric if Eq. (32) holds (with d in place of τ) and d satisfies the triangle inequality,

$$d(x, y) \leqslant d(x, z) + d(z, y), \tag{33}$$

for all x, y, and z. The "almost" metric part for our d to be constructed refers to the fact that the triangle inequality will be guaranteed to hold only for x and y in some arbitrarily chosen compact subset K of X, and for x and y any pre-fixed positive τ distance away from each other, and with z's not too τ-close to x or y. In many applications, these are not restrictive requirements.

In addition we will have the property that $\tau(x, y) \leqslant \tau(u, v) \Longleftrightarrow d(x, y) \leqslant d(u, v)$. As a bonus, a ball with respect to τ will also be a ball with respect to d, and vice versa.

7. A basic example—square of Euclidean metric on $\mathbb{R}^n$

Let us illustrate the various definitions above for an elementary, but still instructive example, namely $X = \mathbb{R}^n$, and $\tau(u, v) = c|u - v|^2$, a scaled square of the Euclidean metric (here $c > 0$ is an overall constant independent of u and v). Then, $N_x(\tau(x, y)) = \{z\colon\ c|x - z|^2 \leqslant c|x - y|^2\} = \{z\colon\ |x - z| \leqslant |x - y|\}$, the ball with center x, radius $|x - y|$, and similarly, $N_y(\tau(x, y))$ is the ball with center y and radius $|x - y|$. The "bad" set $B_{x,y}$ is: $B_{x,y} = \{z\colon\ |x - y|^2 \geqslant |x - z|^2 + |z - y|^2\}$, which is the ball with boundary passing through the points x and y, and whose center is the midpoint of the segment connecting x and y.

In Fig. 5, the two larger balls are $N_x(\tau(x, y))$ and $N_y(\tau(x, y))$, and the eye-shaped region formed by their intersection contains the "pupil" $B_{x,y}$.

8. The construction of a metric from a distance

In this section, we return to the general case of X and distance τ, and show how to construct a(n almost) metric from τ. The restrictions on triplets x, y, and z which we make to ensure that the triangle inequality holds are detailed below.

If $z \in B_{x,y}$, we have that $\tau(x,y) \geqslant \tau(x,z) + \tau(z,y)$, which is equivalent to:

$$1 \geqslant \frac{\tau(x,z)}{\tau(x,y)} + \frac{\tau(z,y)}{\tau(x,y)}. \tag{34}$$

Let

$$\eta_z \equiv \frac{\tau(x,z)}{\tau(x,y)}, \qquad \beta_z \equiv \frac{\tau(z,y)}{\tau(x,y)}. \tag{35}$$

(We will sometimes omit the subscript z.) Note: $\eta_z \geqslant 0$, $\beta_z \geqslant 0$, and $\eta_z + \beta_z \leqslant 1$. Let

$$\widetilde{B}_{x,y} = \big\{(\eta_z, \beta_z)\colon\ z \in B_{x,y}\big\} \subseteq \mathbb{R}^2. \tag{36}$$

Note that $B_{x,y}$ is a subset of X, while $\widetilde{B}_{x,y}$ is by definition always a subset of $\mathbb{R}^2$. In fact, $\widetilde{B}_{x,y} \subseteq [0,1] \times [0,1]$. Remembering our earlier assumptions, $z \to (\frac{\tau(x,z)}{\tau(x,y)}, \frac{\tau(z,y)}{\tau(x,y)})$ is a continuous mapping from $B_{x,y}$ to $\widetilde{B}_{x,y}$ and, as we showed earlier, $B_{x,y}$ is compact, hence $\widetilde{B}_{x,y}$ is compact as well.

Also note that $(0,\alpha) \notin \widetilde{B}_{x,y}$, and $(\alpha,0) \notin \widetilde{B}_{x,y}$ for any $0 \leqslant \alpha < 1$. This can be easily seen since if, say, $(0,\alpha) \in \widetilde{B}_{x,y}$, we see that $\tau(x,z) = 0$, which implies that $x = z$ (by our starting assumptions), which in turn implies that $\alpha = \beta_z = \frac{\tau(z,y)}{\tau(x,y)} = 1$.

Now, let P_η denote projection on the η axis. For every η such that $P_\eta(\widetilde{B}_{x,y}) \neq \emptyset$, let $\phi(\eta) = \inf\{\beta\colon\ (\eta,\beta) \in \widetilde{B}_{x,y}\} = \min\{\beta\colon\ (\eta,\beta) \in \widetilde{B}_{x,y}\}$, with the second equality using the fact that, since $\widetilde{B}_{x,y}$ is compact, so is $\{\beta\colon\ (\eta,\beta) \in \widetilde{B}_{x,y}\}$, for any fixed η. From the discussion just above, we see that if $0 \leqslant \eta < 1$, $\phi(\eta) > 0$.

Without some additional information about how quickly $\phi(\eta)$ converges to 0 as $\eta \to 1$, we cannot, in general say that there is a p with $1 \geqslant p > 0$ such that the set $\widetilde{B}_{x,y}$ lies on or above the graph of $\eta^p + \beta^p = 1$, for all $0 \leqslant \eta \leqslant 1$. For the rest of this note, we fix some α, with $0 < \alpha < 1$, say $\alpha = 0.9$ or 0.99 or 0.999. Then, using the compactness of $\widetilde{B}_{x,y}$, the definition of $\phi(\eta)$ above, and the fact that for $0 \leqslant \eta < 1$, $\phi(\eta) > 0$, there does exist a p with $1 \geqslant p > 0$ such the set $\widetilde{B}_{x,y} \cap ([0,\alpha] \times [0,\alpha])$ lies on or above the graph of $\eta^p + \beta^p = 1$.

Let $p_{x,y}$ be the largest such p (for the particular x and y, with $x \neq y$). Now, vary the pair x, y, and consider the map $\zeta : (x,y) \to p_{x,y}$. By continuity of τ, ζ is continuous at (x,y) if $x \neq y$ (if $x = y$, $p = 1$ works, but as $\tau(x,y) \to 0$, the $p_{x,y}$'s need not be close to 1, and there seems to be no a priori reason that they converge to some specific value). So, we will restrict the domain of ζ as follows. Let $C_\delta = \{(x,y) \in X \times X\colon\ \tau(x,y) \geqslant \delta\}$, where $\delta > 0$ is henceforth fixed and denotes, e.g., the "smallest" distances we want to consider, say $\delta = 1$. The numerical value of δ will, of course, depend on the normalization of the distance, i.e., on the unit of length used. Additionally, we fix a (possibly large) compact subset K of X. Then, $\zeta : C_\delta \cap (K \times K) \to (0,1]$ given by $\zeta(x,y) = p_{x,y}$ is a continuous function on a compact set, so achieves its minimum value, call it $p_{\min}$, and $p_{\min} > 0$.

We now define

$$d(x,y) = \big[\tau(x,y)\big]^{p_{\min}}. \tag{37}$$

Take an arbitrary $(x,y) \in C_\delta \cap (K \times K)$, and select any z. We have two cases, depending on whether z is an element of $B_{x,y}$ or not. If $z \in B_{x,y}$, as discussed above, we will only consider z if it satisfies $\tau(x,z) \leqslant \alpha\tau(x,y)$ and $\tau(y,z) \leqslant \alpha\tau(x,y)$ for some α fixed (close to 1). Then by construction of $p_{x,y}$, we have that:

$$\left[\frac{\tau(x,z)}{\tau(x,y)}\right]^{p_{x,y}} + \left[\frac{\tau(y,z)}{\tau(x,y)}\right]^{p_{x,y}} \geqslant 1. \tag{38}$$

Since $p_{x,y} \geqslant p_{\min}$ and $\frac{\tau(x,z)}{\tau(x,y)} \leqslant 1$ and $\frac{\tau(y,z)}{\tau(x,y)} \leqslant 1$ (due to $z \in B_{x,y}$), we have a fortiori that

$$\left[\frac{\tau(x,z)}{\tau(x,y)}\right]^{p_{\min}} + \left[\frac{\tau(y,z)}{\tau(x,y)}\right]^{p_{\min}} \geqslant 1, \tag{39}$$

hence

$$\big[\tau(x,z)\big]^{p_{\min}} + \big[\tau(y,z)\big]^{p_{\min}} \geqslant \big[\tau(x,y)\big]^{p_{\min}}, \tag{40}$$

i.e.,

$$d(x,z) + d(y,z) \geqslant d(x,y). \tag{41}$$

If $z \notin B_{x,y}$, then $\tau(x,z) + \tau(y,z) \geqslant \tau(x,y)$, and we have:

$$\big[\tau(x,y)\big]^{p_{\min}} \leqslant \big[\tau(x,z) + \tau(y,z)\big]^{p_{\min}} \leqslant \big[\tau(x,z)\big]^{p_{\min}} + \big[\tau(y,z)\big]^{p_{\min}}, \tag{42}$$

again leading to $d(x,y) \leqslant d(x,z) + d(z,y)$. (We have used the inequality $(a+b)^p \leqslant a^p + b^p$, for $a, b \geqslant 0$, and $0 \leqslant p \leqslant 1$.)

We have thus shown that $d(\cdot,\cdot)$ satisfies the triangle inequality (with the restrictions on x, y, and z noted in the above discussion).

The above discussion establishes the following theorem:

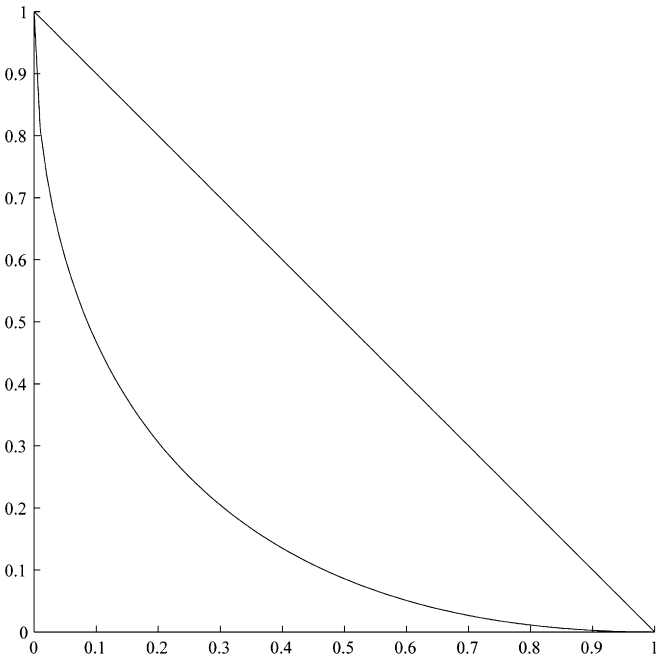

Fig. 6. The set $\widetilde{B}_{x,y}$.

Theorem 8.1. *Let X be a topological space with the property that compact sets are closed. Let $\tau : X \times X \to [0,\infty)$ be a continuous map for which the following properties hold*:

$$\tau(x,y) = 0 \quad \Longleftrightarrow \quad x = y; \qquad \tau(x,y) = \tau(y,x). \tag{43}$$

Further assume that for every $x \in X$, the function $\tau_x(z) \equiv \tau(x,z) : X \to [0,\infty)$ has the property that $\tau_x^{-1}([0,a])$ is a compact subset of X for all $a \geqslant 0$.

Then, for any α, δ with $0 < \alpha < 1$ and $0 < \delta$, and any compact subset K of X, there exists a positive p such that for $(x,y,z) \in K \times K \times X$ with $\tau(x,z) \leqslant \alpha\tau(x,y)$, $\tau(y,z) \leqslant \alpha\tau(x,y)$ and $\tau(x,y) \geqslant \delta$, we have

$$\big[\tau(x,y)\big]^p \leqslant \big[\tau(x,z)\big]^p + \big[\tau(z,y)\big]^p. \tag{44}$$

For comparison, we draw the reader's attention to Theorem 1.1 in [6] which states that a quasi-norm ($\|x+y\| \leqslant C(\|x\| + \|y\|$ instead of the usual triangle inequality) will be subadditive (satisfy the triangle inequality) when raised to a suitable power $p > 0$. Our result above does not assume a quasi-norm like property for τ and we believe has a more constructive proof than the type of argument given for Theorem 1.1 in [6]. However, τ^p in our theorem satisfies the triangle inequality for triplets x, y, z which have additional constraints placed on them (see above).

9. The basic example continued

We return to our earlier example where $X = \mathbb{R}^n$, and $\tau(u,v) = c|u-v|^2$, a scaled square of the Euclidean metric. Using the notation and definitions of the previous section, we see that:

$$\eta_z = |x-z|^2/|x-y|^2, \qquad \beta_z = |y-z|^2/|x-y|^2. \tag{45}$$

It is easy to see that, for any x and y with $x \neq y$, the set $\widetilde{B}_{x,y}$ is exactly the interior and boundary of the lozenge shaped region in Fig. 6, where the graph of the lower boundary is given by: $\sqrt{\eta} + \sqrt{\beta} = 1$. Hence, $p_{x,y} = 1/2$ for all $x \neq y$, and so $p_{\min} = 1/2$. For this example, there is no need for a restricting α, no need for a minimum δ, and no need to intersect with a compact set K. The metric $d(\cdot,\cdot)$ we obtain is then the usual Euclidean metric, scaled by an overall constant.

More generally, if $X = \mathbb{R}^n$, and $\tau(u,v) = c|u-v|^r$, where $r \geqslant 1$, then $p_{\min} = 1/r$, and we obtain the usual Euclidean distance (scaled by an overall constant).

Note that the following hold for our construction:

(1) Since for $x \in X$, $\{y\colon\ [\tau(x,y)]^{p_{\min}} < c\} = \{y\colon\ \tau(x,y) < c^{1/p_{\min}}\}$, balls with respect to τ are also balls with respect to d.
(2) $\tau(x,y) \leqslant \tau(u,v) \Longleftrightarrow d(x,y) \leqslant d(u,v)$.
(3) If τ is already a metric, our construction will yield $p_{\min} = 1$, and $d(\cdot,\cdot) = \tau(\cdot,\cdot)$.

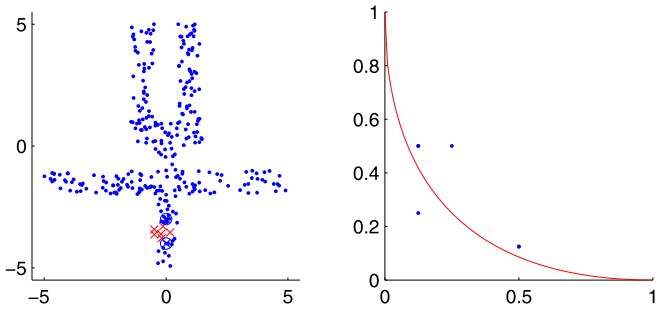

Fig. 7. The "bad" set for points $(0,-4)$ and $(0,-3)$ and the corresponding (η,β) plot.

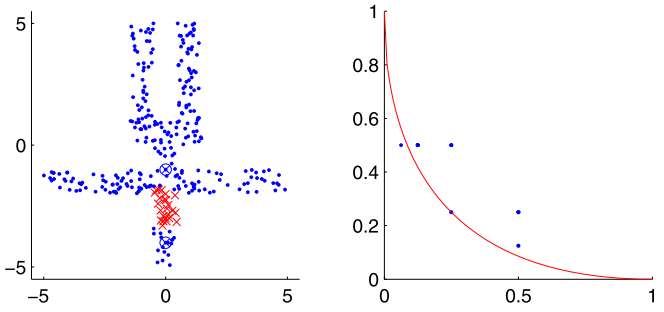

Fig. 8. The "bad" set for points $(0,-4)$ and $(0,-1)$ and the corresponding (η,β) plot.

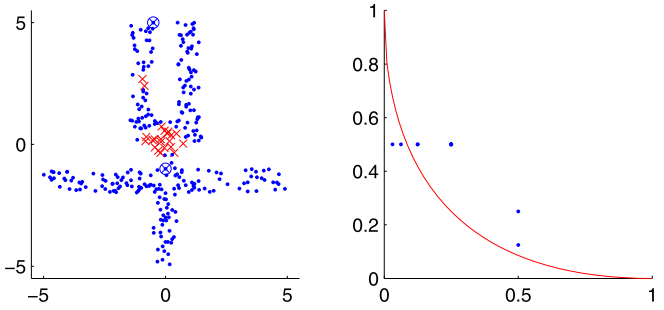

Fig. 9. The "bad" set for points $(0,-1)$ and $(-1/2,5)$ and the corresponding (η,β) plot.

10. The flower set revisited

In this section, we return to our synthetic data set—the flower introduced in Section 5—to illustrate the constructs in Section 8. Each of Figs. 7, 8, and 9 shows in the left plots the "bad" set $B_{x,y}$ for x and y, successively, the pairs $(0,-4)$ and $(0,-3)$; $(0,-4)$ and $(0,-1)$; and $(0,-1)$ and $(-1/2,5)$. The right-hand plot in each figure shows the corresponding (η,β) plot of $\widetilde{B}_{x,y}$; the curve in each of the latter plots is the graph of $\eta^{1/2}+\beta^{1/2}=1$. In the case of heat flow on $\mathbb{R}^2$, all the points of $\widetilde{B}_{x,y}$, for all x and y, lie on or above this curve; see Section 13.

We see in the left-hand plot of Fig. 7 that the "bad" set for the pair $(0,-4)$ and $(0,-3)$ is roughly located inside a circle which has the points $(0,-4)$ and $(0,-3)$ as the endpoints of a diameter; this circle is exactly the "bad" set region for the case of heat flow on $\mathbb{R}^2$, see Section 13. Our random walk for the flower set seems to be behaving in roughly the same way as heat flow on $\mathbb{R}^2$, in this region of the flower set which "looks" like $\mathbb{R}^2$. In Figs. 8 and 9, as the respective pairs of points lie in regions which (at the scale of a Euclidean ball containing each pair of points) look less and less like $\mathbb{R}^2$, the "bad" sets differ more and more markedly from the case of heat flow on $\mathbb{R}^2$. In our experiments, we additionally observed that the pair of points $(-1/2,5)$ and $(1/2,5)$ did not have any associated "bad" points at all.

We also note that each (η,β) plot shows only a few distinct points: since the values of τ_{approx} are dyadic integers, the ratios used as points in the (η,β) plots have a very limited number of possible values.

To obtain the plot shown in Fig. 10, for each value of q $(2\leqslant q\leqslant 4)$ spaced 0.1 apart, we calculated the ratio of points from the "bad" sets arising from *all* pairs of distinct points in the flower set which lie on or above the curve $\eta^{1/q}+\beta^{1/q}=1$ in the (η,β) plane, to the total number of bad points. As can be seen from Fig. 10, letting $p_{\min}=1/2.8$ we see that more

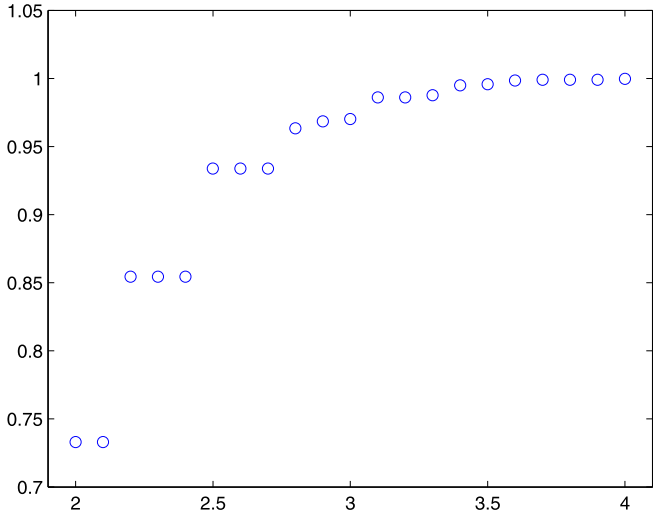

Fig. 10. Ratio of "bad" points above the curve $\eta^{1/q} + \beta^{1/q} = 1$, q on horizontal axis.

than 95% of the points which violate the triangle inequality no longer do so if we use $(\tau_{approx})^{1/2.8}$ instead of τ_{approx}. As will be shown in Section 13 below, for heat flow on $\mathbb{R}^n$, $p_{\min} = 1/2$.

11. Constructively approximating $p_{\min}$

In this section, we give an informal sketch of a proposed way to efficiently approximate $p_{\min}$. Our method may yield a value for $p_{\min}$ that is higher than the true value, but has the advantage of being efficient and checking across scales. We can improve the value of the exponent if we have more computational resources at our disposal and can check more points; the earlier computations can be "recycled" and need not be done over again.

Let us suppose that the smallest distances in which we will be interested are of magnitude 1, and that we will only be calculating distances between points in some compact $K \subseteq X$. Consider a (finite) list of points $x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, \ldots \in K$ such that $\tau(x_i^{(0)}, x_j^{(0)}) \approx 1$, if $i \neq j$, and, for every $y \in K$, there is some $x_i^{(0)}$ with $\tau(x_i^{(0)}, y)$ not much larger than 1. It does not matter if the sets $\{y\colon\ \tau(x_i^{(0)}, y) < 1\}_i$ overlap. For each $x_i^{(0)}$, look at the nearest $x_j^{(0)}$'s in the list, say $x_{i_1}^{(0)}, x_{i_2}^{(0)}, \ldots, x_{i_k}^{(0)}$. For each of these $x_{i_j}^{(0)}$, sample $B_{x_i^{(0)}, x_{i_j}^{(0)}}$: since $B_{x_i^{(0)}, x_{i_j}^{(0)}}$ is compact and contained in the intersection of $N_{x_i^{(0)}}(\tau(x_i^{(0)}, x_{i_j}^{(0)}))$ and $N_{x_{i_j}^{(0)}}(\tau(x_i^{(0)}, x_{i_j}^{(0)}))$, we are searching locally. Next, plot the (η, β) points in $\widetilde{B}_{x_i^{(0)}, x_{i_j}^{(0)}}$ as discussed in Section 8; plot all the (η, β) points for all the $x_{i_1}^{(0)}, x_{i_2}^{(0)}, \ldots, x_{i_k}^{(0)}$ on the *same* η, β plane.

Continue for the other $x_i^{(0)}$'s. If we come to a point $x_i^{(0)}$ where the points (η, β) are below the points already plotted, sample a few more points nearby with τ distance about 1 from the current point $x_i^{(0)}$, thus adaptively oversampling in areas where something is happening in X that causes the (η, β) points to dip more than "usual".

A Monte-Carlo variation on the above process is to just pick many random pairs $(x_i^{(0)}, x_j^{(0)})$ in K, and not worry about some parts of K being farther than distance 1 away from any of the points selected.

Next, either subsample the $x_i^{(0)}$'s, or generate new points, to get a finite list $x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \ldots \in K$ so that $\tau(x_i^{(1)}, x_j^{(1)}) \approx 2$, if $i \neq j$. Repeat the above procedure of plotting the pairs (η, β), for each $x_i^{(1)}$ and its nearest neighbors in the list, still on the *same* η, β plane.

Keep going: at the nth stage, get a finite list $x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, \ldots \in K$ so that $\tau(x_i^{(n)}, x_j^{(n)}) \approx 2^n$, if $i \neq j$. Keep plotting the pairs (η, β), for each $x_i^{(n)}$ and its nearest neighbors in the list, on the original η, β plane.

Finally, select the greatest $p_{\min}$, with $0 < p_{\min} \leqslant 1$ such that all, or all except some outliers, of the plotted (η, β) points are located on or above the curve $\eta^{p_{\min}} + \beta^{p_{\min}} = 1$, in the square $[0, \alpha] \times [0, \alpha]$ in the η, β plane.

Note that the $p_{\min}$ we have constructed is not less, but could be greater, than the "true" $p_{\min}$ since we sampled only some points in the "bad" sets for some pairs of points in K, across a particular scale ladder. However, finding our $p_{\min}$, and defining $d(x, y) = [\tau(x, y)]^{p_{\min}}$, we have deformed the original τ across many pairs of points scattered over K, and at different τ scales, so should not be too far from the "true" $p_{\min}$.

We finish this section by summarizing the above discussion by an "informal" pseudocode.

The input is a (finite) discrete set X with a distance function τ, a number α, $0 < \alpha < 1$, and the number of levels L. We also assume we have generated a set of points $\{x_1, x_2, x_3, \ldots, x_n\}$ in X such that $\tau(x_i, x_j) \approx 1$ if $i \neq j$ and, for any $y \in X$

there is some x_i with $\tau(x_i, y)$ not much larger than 1. We suppose that the cost to calculate the τ-distance between any two points is bounded by a constant.

The output is an exponent p so that τ^p satisfies the triangle inequality for various pairs of points of the input sequence $\{x_1, x_2, x_3, \ldots, x_n\}$, across different scales (see pseudocode):

```
create (η, β) plot P, no points plotted yet;
dist := 1;
current set = {x_1, x_2, x_3, ..., x_n};
while dist ⩽ 2^L and current set has at least 2 points
    loop over the elements x of current set
        find all y in current set such that τ(x, y) ⩽ 2dist;
        loop over these y
            plot all pairs of B̃_{x,y} on the plot P;
        end loop
    end loop
    subsample current set to get new set where points are roughly
        2dist apart and every point of current set is not too far
        from a point in new set;
    current set := new set;
    dist = 2dist;
end while
p := greatest number between 0 and 1 such that all
    (or all except some outliers) of the plotted
    (η, β) points are located on or above the curve η^p + β^p = 1,
    in the square [0, α] × [0, α] on the plot P;
```

Let M be the maximum of the computation costs of executing "find all y in *current set* such that $\tau(x, y) \leqslant 2dist$"; note that the neighborhood size keeps growing with *dist* while the separation between points of *current set* is increasing at the same rate.

We see that the computational cost of the above procedure is bounded by $CLMn$, where C is a constant.

12. Making a metric out of the distances τ_J or τ_{approx}

Returning to the first part of the paper, we can apply the above construction of a(n almost) metric to Jones's distance τ_J, see Section 3 or τ_{approx}, see Section 4. If our data set S is not too large, we can skip the tree construction entirely, and just use the second part of this paper to obtain a(n approximate) metric from τ_J. If S is large however, using τ_J as a distance between any two pairs of points would be computationally expensive, and we could use the tree construction discussed in the first part of the paper to find τ_{approx} efficiently for any pair of points, and then use the construction in the second part to obtain a(n approximate) metric from τ_{approx}. In the latter case, the proposed procedure in Section 11 above could be used to approximate the value of $p_{\min}$. Since the sets $S_j = \{y_1^{(j)}, y_2^{(j)}, \ldots, y_{i_j}^{(j)}\}$ introduced in Section 4 serve the purpose of sampling the data at coarser and coarser grids in a closely related fashion to increasingly coarser τ_{approx} separation, we suggest that the sets S_j themselves be used as the points $x_1^{(j)}, x_2^{(j)}, x_3^{(j)}, \ldots$ in Section 11.

13. Application to heat flow on $\mathbb{R}^n$

In the case of heat flow on $X = \mathbb{R}^n$, we have that:

$$\rho_t(x, y) = (4\pi t)^{-n/2} e^{-|x-y|^2/4t}. \tag{46}$$

Using Eq. (60) in Appendix A, we see that:

$$\frac{1}{2}\left\|\rho_t(\cdot, x) - \rho_t(\cdot, y)\right\|_1 = F\left(\frac{|x-y|}{2\sqrt{t}}\right), \tag{47}$$

where

$$F(\beta) \equiv c \int_0^\beta e^{-s^2/4}\, ds, \tag{48}$$

and c is such that $F(\infty) = 1$.

With $\tau_J(\cdot,\cdot)$ as in Section 3, using the third line of (14) (as well as Proposition 3.2), and since the function F is strictly increasing, we see that

$$\sqrt{\tau_J(x,y)} = \frac{|x-y|}{2F^{-1}(1-\gamma)}, \tag{49}$$

hence

$$\tau_J(x,y) = C_\gamma |x-y|^2, \tag{50}$$

where C_γ is an overall constant depending only on γ. But then $\tau_J(\cdot,\cdot)$ is exactly the distance considered in our "Example" sections, Sections 7 and 9, and leads to (scaled) Euclidean metric by using our construction of $d(\cdot,\cdot)$ in Section 8. Note that $0<\gamma<1$ is a freely chosen parameter: however, the metric $d(\cdot,\cdot)$ that results is fundamentally independent of the choice of γ since that choice is only reflected in the value of an overall scaling.

14. Additional remarks

We begin this section by discussing a distance similar to that proposed by Peter Jones (see Section 3, in particular Eq. (14)), but one which uses a measure of separation between diffusion "bumps" other than L^1. This alternate separation is related to cosine similarity, and is the length of the chord joining the tips, on the unit hypersphere, of two L^2 normalized vectors. (The authors of this paper have discussed some advantages of using cosine similarity for measuring diffusion separation in [8].) In the second part of this section, we present some indications that at least for this alternative to Jones's distance, the construction we have described in Section 8 should work well for more general situations than just heat flow on $\mathbb{R}^n$.

Let us consider again Eq. (14) in Section 3. We do not have to use L^1 to measure separation between diffusion bumps; for example, for a fixed parameter γ with $0<\gamma<1$, we may define a distance $\tau_{\text{alt}}(x,y)$ by:

$$\begin{aligned}\tau_{\text{alt}}(x,y) &\equiv \operatorname{arginf}_t\left\{\frac{1}{\sqrt{2}}\left\|\frac{\rho_t(\cdot,x)}{\|\rho_t(\cdot,x)\|_2}-\frac{\rho_t(\cdot,y)}{\|\rho_t(\cdot,y)\|_2}\right\|_2 \leqslant 1-\gamma\right\}\\ &= \operatorname{arginf}_t\left\{\left(1-\frac{\int_X \rho_t(u,x)\rho_t(u,y)\,du}{\|\rho_t(\cdot,x)\|_2\|\rho_t(\cdot,y)\|_2}\right)^{1/2}\leqslant 1-\gamma\right\},\end{aligned} \tag{51}$$

where

$$\left\|\rho_t(\cdot,z)\right\|_2 = \left(\int_X \rho_t(u,z)\rho_t(u,z)\,du\right)^{1/2}. \tag{52}$$

If the semi-group kernel $\rho_t(\cdot,\cdot)$ is symmetric, the advantage of using $\tau_{\text{alt}}(x,y)$ instead of $\tau_J(x,y)$ is that the integrals that appear in the definition of the former can be evaluated by using the semi-group property (1):

$$\tau_{\text{alt}}(x,y) = \operatorname{arginf}_t\left\{\left(1-\frac{\rho_{2t}(x,y)}{\sqrt{\rho_{2t}(x,x)}\sqrt{\rho_{2t}(y,y)}}\right)^{1/2}\leqslant 1-\gamma\right\}. \tag{53}$$

Unlike the L^1 norm, appearing in the definition of $\tau_J(x,y)$, which decreases with t (see Corollary 3.1), we do not know whether the L^2 norm of the difference, appearing in the definition of $\tau_{\text{alt}}(x,y)$, is in general monotonic in t, even in the case of a symmetric $\rho_t(\cdot,\cdot)$.

For the case of heat flow on $X=\mathbb{R}^n$, calculating $\tau_{\text{alt}}(x,y)$ (note: the L^2 norm of the difference *is* monotonic in t for heat flow in $\mathbb{R}^n$) leads to:

$$\tau_{\text{alt}}(x,y) = \frac{|x-y|^2}{8\ln(1/(1-(1-\gamma)^2))} \equiv \widetilde{C}_\gamma |x-y|^2, \tag{54}$$

where $\widetilde{C}_\gamma$ is again an overall constant depending only on γ. Hence our distance construction in Section 8, starting with $\tau_{\text{alt}}(\cdot,\cdot)$, instead of $\tau_J(\cdot,\cdot)$, again leads to (scaled) Euclidean distance as in Section 13.

The construction we have described in Section 8 defines a metric (actually, an "almost" metric) by $d(x,y)=[\tau(x,y)]^{p_{\min}}$, where $0<p_{\min}\leqslant 1$, for a given distance τ. We have described a theoretical way to find $p_{\min}$, and a numerical approach to find an approximation to $p_{\min}$. Of course, if we take $p_{\min}$ to be too small, we stop being able to distinguish different distances: if we take the extreme case of $p_{\min}=0$, the triangle inequality would be saying that $1\leqslant 1+1$, a not very useful statement. Thus a natural question is whether in practical situations we can find a $p_{\min}$ that is far enough away from 0 to be useful.

We have one case in which the construction we have described works well: the case of heat flow on $\mathbb{R}^n$ in which we obtain exactly the right metric, starting from both τ_J and τ_{alt}. What about more general cases? There are some indications in the literature that the construction we have described in Section 8 should work well for more general situations.

For example, the author of [3, Chapter XII, Section 12] refers to the following results: If M is a complete Riemannian manifold with Ricci curvature bounded below by $(n-1)\kappa$, $\kappa \leqslant 0$, and $\rho_t(x, y)$ is the heat kernel on M, then (with $V(z; s)$ the volume of the ball centered at z with geodesic radius s and $d(x, y)$ the geodesic distance (metric), in the following formulas):

$$\rho_t(x, y) \leqslant \frac{c(n, \delta)}{V^{1/2}(B(x; \sqrt{t}))V^{1/2}(B(y; \sqrt{t}))} \exp\left\{-\frac{d^2(x, y)}{(4+\delta)t} - c(n)\delta\kappa t\right\}, \tag{55}$$

and

$$\rho_t(x, y) \geqslant \frac{c(n, \epsilon)}{V^{1/2}(B(x; \sqrt{t}))V^{1/2}(B(y; \sqrt{t}))} \exp\left\{-\frac{d^2(x, y)}{(4-\epsilon)t} + c(n)\epsilon\kappa t\right\}. \tag{56}$$

We see that if $\kappa = 0$, the estimates are particularly simple and very close to the situation of heat flow in $\mathbb{R}^n$. The boundedness results above are developed in [12].

In the situation just described, using $\tau_{\text{alt}}(x, y)$ as the distance for the case above (and the fact that the heat kernel on a manifold is symmetric in the two space arguments), Eq. (53) suggests that our construction would yield $p_{\min}$ of the "order" of $1/2$, and the metric we will arrive at would have a close relation to the geodesic metric on M.

As another example, the authors of [1] obtain sub-Gaussian heat kernel upper estimates for certain weighted graphs, of the form:

$$\frac{C}{V(x, n^{1/\beta})} \exp\left(-\left(\frac{d(x, y)^\beta}{Cn}\right)^{\frac{1}{\beta-1}}\right). \tag{57}$$

Here, x and y are vertices in the graph, $\beta \geqslant 2$, and n is any natural number and denotes the number of steps taken in a random walk. Similar estimates hold for lower bounds.

In this case as well, using $\tau_{\text{alt}}(x, y)$ as the distance, with n playing the role of t, suggests that our construction would yield $p_{\min}$ of the "order" of $1/\beta$, and the metric we will arrive at would have a close relation to the shortest vertex distance d on the graph.

15. Conclusions

In this paper, we have discussed two different, yet related, constructions. In the first part of the paper, we have presented a tree-based approach to compute an approximation to Jones's diffusion distance.

The original diffusion distance proposed by Coifman, when applied to heat flow on $\mathbb{R}^n$, gives Euclidean distance only locally, with a constant which depends both on the time t chosen and the dimension n; see Appendix B. Moreover, as discussed in Appendix B, using L^2 as was originally proposed by Coifman has some scaling issues: the probability density functions which are used are unit vectors in L^1, not L^2. Jones's distance, when applied to heat flow on $\mathbb{R}^n$, gives the scaled square of the Euclidean distance globally, as well as with a constant independent of the dimension n. Since in this most basic example of heat flow recovering a dimension-independent Euclidean distance seems intuitively desirable, we believe Jones's distance is worthy of consideration. Furthermore, Jones's proposed distance can be used for the case of non-symmetric diffusion (such as say a directed wind blowing across a spreading fire) without any modification. Moreover, considering the first time that probability densities from two points are close enough to each other is, to us, an appealing concept worthy of further examination on aesthetic grounds.

Computing Jones's distance directly for the case of a discrete data set with a transition matrix involves considering increasing powers of the transition matrix and the l^1 differences of two rows of these powers corresponding to the two points involved. If the data set is large, taking powers of the transition matrix is expensive. In the first part of our paper, we describe a tree-based construction to compute an approximation to Jones's distance. Assuming that finding neighbors in subsets of the data which consist of points separated at roughly the same level as the neighborhood size is not computationally expensive, we see that our construction yields a more efficient computation. In our proposal, stopping when ancestors of two points are neighbors, rather than being equal, has the effect of avoiding grid artifacts in defining the approximate distance and the necessity of averaging over different realizations of ancestral sequences (which would lead to complicated probabilistic arguments). We then show that our approximation to Jones's distance is comparable to Jones's distance from above and from below.

As mentioned for the case of heat flow on $\mathbb{R}^n$, Jones's distance is not a metric: it fails to satisfy the triangle inequality. Our proposed approximation is not a metric either. Since the triangle inequality seems so natural when discussing distances (after all, Euclidean distance seems preferable to the square of Euclidean distance in $\mathbb{R}^n$) the second part of our paper is a general procedure to construct an "almost" metric from any distance. We give a theoretical description of our method and an informal sketch how one can apply our approach to "metrize" a given distance across different scales. The more computational resources we have, the more points we can include in the metrization procedure. Using more points can only serve to decrease the exponent to which we need to raise distances to satisfy the triangle inequality; the earlier work with previous pairs of points is not "lost" and need not be redone.

Acknowledgments

We are grateful to Raphy Coifman for his continued willingness to discuss mathematics with us. We would also like to thank the reviewers for their helpful comments.

Appendix A

We present some calculations whose results are used in Section 13. First, we calculate

$$M \equiv \int_{-\infty}^{\infty} \left| e^{-(x-a)^2/4} - e^{-x^2/4} \right| dx, \quad a \geqslant 0. \tag{58}$$

Breaking up the above integral, and removing the absolute value bars, we obtain:

$$\begin{aligned}
M &= \int_{a/2}^{\infty} \left(e^{-(x-a)^2/4} - e^{-x^2/4} \right) dx + \int_{-\infty}^{a/2} \left(e^{-x^2/4} - e^{-(x-a)^2/4} \right) dx \\
&= \int_{a/2}^{\infty} e^{-(x-a)^2/4}\, dx - \int_{a/2}^{\infty} e^{-x^2/4}\, dx + \int_{-\infty}^{a/2} e^{-x^2/4}\, dx - \int_{-\infty}^{a/2} e^{-(x-a)^2/4}\, dx \\
&= \int_{-a/2}^{\infty} e^{-x^2/4}\, dx - \int_{a/2}^{\infty} e^{-x^2/4}\, dx + \int_{-\infty}^{a/2} e^{-x^2/4}\, dx - \int_{-\infty}^{-a/2} e^{-x^2/4}\, dx \\
&= 2 \int_{-a/2}^{a/2} e^{-x^2/4}\, dx \\
&= 4 \int_{0}^{a/2} e^{-x^2/4}\, dx.
\end{aligned} \tag{59}$$

Now, for $x, y \in \mathbb{R}^n$ and $t > 0$, we have, with the constant c taking different values from expression to expression as necessary:

$$\begin{aligned}
\frac{c}{t^{n/2}} \int_{\mathbb{R}^n} \left| e^{-|x-u|^2/4t} - e^{-|y-u|^2/4t} \right| du &= \frac{c}{t^{n/2}} \int_{\mathbb{R}^n} \left| e^{-|x-y+v|^2/4t} - e^{-|v|^2/4t} \right| dv \\
&= c \int_{\mathbb{R}^n} \left| e^{-|x-y-w\sqrt{t}|^2/4t} - e^{-|w|^2/4} \right| dw \\
&= c \int_{\mathbb{R}^n} \left| e^{-|(x-y)/\sqrt{t}-w|^2/4} - e^{-|w|^2/4} \right| dw \\
&= c \int_{-\infty}^{\infty} \left| e^{-(|x-y|/\sqrt{t}-s)^2/4} - e^{-s^2/4} \right| ds \\
&= c \int_{0}^{|x-y|/2\sqrt{t}} e^{-s^2/4}\, ds \\
&= F\left(\frac{|x-y|}{2\sqrt{t}} \right),
\end{aligned} \tag{60}$$

where

$$F(\beta) \equiv c \int_{0}^{\beta} e^{-s^2/4}\, ds. \tag{61}$$

In the above derivation, the first equality follows by the change of variables $v = y - u$, the second equality follows from using the change of variables $v = -w\sqrt{t}$, the fourth equality follows by integrating out over the directions orthogonal to the direction of the vector $x - y$, and the fifth equality follows from Eq. (59) above.

Appendix B

We present an expanded discussion of Coifman's original diffusion distance, as well as what we see are some of its drawbacks. For more details and additional discussion see [8]. For notation, please see Section 2.

For the continuous case, the unweighted version of Coifman's distance between $x, y \in X$, which we will denote by $d_{C,t}(x, y)$, can be defined as follows:

$$\begin{aligned}\big[d_{C,t}(x,y)\big]^2 &\equiv \big\langle T_t(\delta_x - \delta_y), T_t(\delta_x - \delta_y)\big\rangle \\ &= \langle T_t\delta_x, T_t\delta_x\rangle + \langle T_t\delta_y, T_t\delta_y\rangle - 2\langle T_t\delta_x, T_t\delta_y\rangle. \end{aligned} \tag{62}$$

The $\langle\,,\,\rangle$ is the usual inner product on $L^2(X)$. (In [4], the authors consider a weighted version of Eq. (62) which naturally arises when the underlying kernel does not integrate to 1 (in each variable). In terms of data analysis, this corresponds to cases where the data are sampled non-uniformly over the region of interest. For simplicity, we are just using Coifman's unweighted distance.)

Note that we thus have:

$$\begin{aligned}\big[d_{C,t}(x,y)\big]^2 &= \int_X \rho_t(v,x)\rho_t(v,x)\,dv + \int_X \rho_t(v,y)\rho_t(v,y)\,dv - 2\int_X \rho_t(v,x)\rho_t(v,y)\,dv \\ &= \big\|\rho_t(\cdot,x)\big\|_2^2 + \big\|\rho_t(\cdot,y)\big\|_2^2 - 2\big\langle\rho_t(\cdot,x), \rho_t(\cdot,y)\big\rangle. \end{aligned} \tag{63}$$

Although Coifman's original definition used a kernel symmetric with respect to the space variable, $d_{C,t}(x, y)$ as given above need not be based on a symmetric ρ_t. Note that, by the defining Eq. (62), $d_{C,t}(x, y)$ is symmetric in x and y (even if ρ_t is not), and satisfies the triangle inequality. If ρ_t is symmetric in the space variables, from Eq. (1) we see that:

$$\big[d_{C,t}(x,y)\big]^2 = \rho_{2t}(x,x) + \rho_{2t}(y,y) - 2\rho_{2t}(x,y), \tag{64}$$

a form matching one of Coifman's formulations for the continuous case.

If, in addition to ρ_t being symmetric in the space variables, we assume (as in the case of heat flow on a compact manifold) that there exist $0 \leqslant \lambda_1 \leqslant \lambda_2 \leqslant \cdots$, with each λ_j corresponding to a finite dimensional eigenspace, and a complete orthonormal family of L^2 functions $\phi_1, \phi_2, \ldots$, such that:

$$\rho_t(x,y) = \sum_{j=1}^{\infty} e^{-\lambda_j t}\phi_j(x)\phi_j(y), \tag{65}$$

for $t > 0$, we easily see that:

$$\big[d_{C,t}(x,y)\big]^2 = \sum_{j=1}^{\infty} e^{-2\lambda_j t}\big(\phi_j(x) - \phi_j(y)\big)^2, \tag{66}$$

the original form proposed by Coifman. Note that the latter expression again explicitly shows that $d_{C,t}(x, y)$ is symmetric in x and y, and satisfies the triangle inequality (by considering, for example, the right-hand side as the square of a weighted distance in l^2).

An important benefit of introducing a diffusion distance as above can be illustrated by considering Eq. (66). If ρ_t is such that Eq. (66) holds for a complete orthonormal family $\{\phi_j\}$, we see that as t increases, we are achieving an (approximate) isometric embedding of X into successively lower dimensional vector spaces (with a weighted norm). More specifically, for $\lambda_j > 0$, if t is large, the terms $e^{-2\lambda_j t}(\phi_j(x) - \phi_j(y))^2$ are nearly 0. So, as t increases, we see that the "heat smeared" X is parameterized by only a few leading ϕ_j's. Thus, "stepping" through higher and higher times, we are obtaining a natural near-parameterization of more and more smeared versions of X, giving rise to a natural ladder of approximations to X.

See [4,10], and [5] for more discussion and examples of the natural embedding discussed above, along with illustrations of its power to organize unordered data, as well as its insensitivity to noise.

We would now like to point out what we see as some drawbacks of Coifman's distance.

Let's consider Eq. (64) for the case where

$$\rho_t(x,y) = (4\pi t)^{-n/2} e^{-|x-y|^2/4t}, \tag{67}$$

the fundamental solution to the heat equation in $\mathbb{R}^n$. Then,

$$\left[d_{C,t}(x,y)\right]^2 = \frac{2-2e^{-|x-y|^2/8t}}{(8\pi t)^{n/2}}. \tag{68}$$

If $|x-y|^2/8t$ is small, then to leading order in $|x-y|^2/4t$,

$$\left[d_{C,t}(x,y)\right]^2 = \frac{1}{(8\pi t)^{n/2}}\left(\frac{|x-y|^2}{4t} + \mathcal{O}\left(\left(\frac{|x-y|^2}{4t}\right)^2\right)\right). \tag{69}$$

Thus, if $|x-y| \ll \sqrt{t}$, we do recover the geodesic distance between x and y (as would be reasonable to expect) but, due to the $1/t^{n/2}$ term in front, normalized by a power of t which depends on the dimension n. For $\mathbb{R}^n$ itself, the normalization does depend on n, but is simply a global change of scale, for each t, and thus basically immaterial. Suppose, however, that the data we are considering come in two "clumps", one of dimension n, and the other of dimension m, with $n \neq m$. Let's also suppose these clumps are somehow joined together and, far away from the joining region, each clump is basically flat Euclidean space of the corresponding dimension. Then, far away from the joint, heat diffusion in a particular clump would behave as if it were in $\mathbb{R}^n$, respectively $\mathbb{R}^m$ (until the time that the flowing heat "hits" the joint region). Thus, in the part of each clump that is far from the joint, the diffusion distance would be normalized differently, one normalization depending on n and the other on m. An overall change of scale would not remove this difference, thus we would not recover the usual Euclidean distance in the two clumps simultaneously, as we would like.

The second point of concern is more general in nature. In the continuous case, Coifman's distance involves the L^2 distance between $T_t\delta_z$, when $z = x$, and $T_t\delta_z$ when $z = y$, see Eq. (62). The L^1 norm of $T_t\delta_z$ is 1, using the mass preservation assumption of Eq. (3). So the diffusion distance proposed by Coifman finds the L^2 distance between L^1 normalized vectors.

Let's illustrate by an example, in the discrete situation for variety, in which this may lead to undesired results. Let $N = 10{,}000$ and, without specifying the matrix A, suppose that after some time has passed, we have the following two $1 \times 10{,}000$ vectors giving two different results of diffusion:

$$v_1 = \left(\frac{1}{100}, \ldots, \frac{1}{100}, 0, \ldots, 0\right),$$

where the first one hundred entries are each $1/100$, and the rest 9900 entries are 0, and

$$v_2 = \left(\frac{1}{10{,}000}, \frac{1}{10{,}000}, \ldots, \frac{1}{10{,}000}\right),$$

where each entry is $1/10{,}000$.

Note that v_1 and v_2 both have l^1 norm 1. Now, considering two canonical basis vectors e_i^T and e_j^T, $i \neq j$, each of which has l^1 norm 1, we see that $\langle e_i^T - e_j^T, e_i^T - e_j^T\rangle = 2$. So, a distance of $\sqrt{2}$ gives the (in fact, maximum) separation between two completely different (l^1 unit) diffusion vectors. Return to v_1 and v_2, note that v_2 corresponds to total diffusion, while v_1 has only diffused over 1% of the entries. We would thus hope that v_1 and v_2 would be nearly as much separated as e_i^T and e_j^T, i.e. have diffusion distance not much smaller than $\sqrt{2}$. But a trivial calculation shows that:

$$\sqrt{\langle v_1 - v_2, v_1 - v_2\rangle} < 0.1,$$

which seems much smaller than what we would like. The problem is that $\sqrt{\langle v_1 - v_2, v_1 - v_2\rangle}$ is small since the l^2 norm of each of v_1 and v_2 is small, even though the l^1 norm of each is 1.

References

[1] M. Barlow, T. Coulhoun, T. Kumagai, Characterization of sub-Gaussian heat kernel estimates on strongly recurrent graphs, Comm. Pure Appl. Math. 58 (12) (2005) 1642–1677.
[2] S.-H. Cha, Comprehensive survey on distance/similarity measures between probability density functions, Int. J. Math. Models Methods Appl. Sci. 1 (4) (2007) 300–307.
[3] I. Chavel, Eigenvalues in Riemannian Geometry, Pure Appl. Math., vol. 115, Academic Press, Orlando, FL, 1984.
[4] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (2006) 5–30.
[5] R.R. Coifman, M. Maggioni, Diffusion wavelets, Appl. Comput. Harmon. Anal. 21 (1) (July 2006) 53–94.
[6] R.A. DeVore, G.G. Lorentz, Constructive Approximation, Springer-Verlag, Berlin, 1993.
[7] E. Deza, M.-M. Deza, Dictionary of Distances, Elsevier, Amsterdam, 2006.
[8] M. Goldberg, S. Kim, Some remarks on diffusion distances, J. Appl. Math. 2010 (2010), Article ID 464815, 17 pp., doi:10.1155/2010/464815.
[9] R.E. Huff, Existence and uniqueness of fixed-points for semigroups of affine maps, Trans. Amer. Math. Soc. 152 (1) (1970) 99–106.
[10] S. Lafon, Diffusion maps and geometric harmonics, PhD thesis, Yale University, 2004.
[11] D. Levin, Y. Peres, E. Wilmer, Markov Chains and Mixing Times, American Mathematical Society, Providence, RI, 2009.
[12] P. Li, S.-T. Yau, On the parabolic kernel of the Schrödinger operator, Acta Math. 156 (1986) 153–201.
[13] B. Nadler, S. Lafon, R.R. Coifman, I.G. Kevrekidis, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems, Appl. Comput. Harmon. Anal. 21 (2006) 113–127.
[14] B. Yood, On fixed points for semi-groups of linear operators, Proc. Amer. Math. Soc. 2 (2) (1951) 225–233.